CONF-961119--3

PACKAGING AND DISTRIBUTING ECOLOGICAL DATA FROM MULTISITE STUDIES*

R.J. Olson, L.D. Voorhees, J.M. Field, and M.J. Gentry**
Oak Ridge National Laboratory***
Oak Ridge, Tennessee 37831-6407, U.S.A.

RECEIVED
SEP 0 9 1996
OSTI

ABSTRACT

Studies of global change and other regional issues depend on ecological data collected at multiple study areas or sites. An information system model is proposed for compiling diverse data from dispersed sources so that the data are consistent, complete, and readily available. The model includes investigators who collect and analyze field measurements, science teams that synthesize data, a project information system that collates data, a data archive center that distributes data to secondary users, and a master data directory that provides broader searching opportunities. Special attention to format consistency is required, such as units of measure, spatial coordinates, dates, and notation for missing values. Often data may need to be enhanced by estimating missing values, aggregating to common temporal units, or adding other related data such as climatic and soils data. Full documentation, an efficient data distribution mechanism, and an equitable way to acknowledge the original source of data are also required.

1.0 INTRODUCTION

Research and assessments of global change often entail the synthesis and modeling of ecological phenomena over regions and require multidisciplinary data from multiple study areas or sites. Thus, an integrated database comprising historical data, newly collected field data, remotely sensed image data, and GIS coverages is needed. To achieve consistency and completeness, the data must be adequately documented, as well as processed and enhanced. The challenge of combining diverse

---

*Presented at Eco-Informa '96, Lake Buena Vista, Florida, 4-7 November 1996.
**University of Tennessee, Knoxville, Tennessee.
***Research sponsored by the National Aeronautics and Space Administration under Interagency Agreement DOE No. 2013-F044-A1 under Lockheed Martin Energy Research Corp. contract DE-AC05-96OR22464 with the U.S. Department of Energy.

"The submitted manuscript has been authorized by a contractor of the U.S. Government under contract No. DE-AC05-96OR22464. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes."

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

MASTER

A-01 1.tif

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

A-02 2.tif

data for global change research has been studied by a National Research Council review committee (NRC, 1995). On the basis of six case studies, the NRC committee identified 18 barriers to the interfacing of data and developed 10 keys to success. Only two of the ten recommendations dealt with technology while the others addressed cultural aspects of management, human behavior, and marketing associated with developing integrated data resources. The companion challenge of ensuring the availability of ecological data has been studied by the Ecological Society of America (Gross et al., 1995), which identified a growing concern that long-term ecological data are lost after projects are completed because investigators move on to other interests and the data are rarely archived.

An example of a large integrated, intersite project is the First ISLSCP (International Satellite Land Surface Climatology Project) Field Experiment (FIFE), funded by the National Aeronautics and Space Administration (NASA). The FIFE project collected intensive remote-sensing and field data on a prairie site in Kansas in 1987 and 1989 to better understand fluxes between the land surface and the atmosphere and to develop associated remote-sensing methodologies. The project included 29 teams organized in 6 disciplinary groups. An extensive collection of FIFE data and metadata were compiled and distributed as a 5-volume CD-ROM set (Strebel et al., 1994a), and the FIFE information system was used as an example of a conceptual framework for scientific information systems (Strebel et al., 1994b). A recent compilation of 300 publications associated with FIFE and a survey of users of FIFE data highlight the scientific value of readily available, well-documented data to a multidisciplinary study (see the World Wide Web site at http://www-eosdis.ornl.gov/).

Projects conducted at Oak Ridge National Laboratory (ORNL) have employed a process to produce integrated databases for use in modeling and assessments at the regional and global scale. Currently ORNL directs the Distributed Active Archive Center (DAAC) for Biogeochemical Dynamics, as part of NASA's Earth Observing System Data and Information System (EOSDIS) Project. EOSDIS forms an integral part of NASA's contribution to the Global Change Research Program. Nine DAACs plus Affiliated Data Centers archive and distribute EOSDIS data. Drawing upon the ORNL DAAC's experience in collecting and distributing data, this paper proposes an information system model to package and distribute ecological data for studies of global change.

The information system model incorporates data processing as an integral component of ongoing activities associated with a project, especially among science teams within a project or at individual sites of a multisite project. Although the needs for a specific project may provide the incentive and focus for assembling, enhancing, and documenting multisite data, the model proposed here contributes to creating

A-03 3.tif

specific dataset metadata are the EOSDIS Guide document (see Web site at http://harp.gsfc.nasa.gov:1729/eosdis_documents/guideshells/dataset.html/) and a proposed nonspatial metadata standard for ecological data (Michener et al., in press). The GCMD and EOSDIS activities are both developing and maintaining controlled sets of keywords to characterize datasets.

Similar to writing scientific papers, preparing metadata is a joint effort among the author(s), technical editors from the project information system staff, and peer reviewers representing science teams or potential dataset users. Investigators responsible for collecting the data are expected and encouraged to provide a first draft of their dataset metadata for a review process. The project information system staff review the initial metadata in terms of format, completeness, and consistency, including adding standard publication elements such as keywords and acronym definitions. Review comments are evaluated by the project information system staff and, if revisions are indicated, the author. The project information system group will incorporate changes based on the author's approval. Finally, the metadata will accompany the data and be submitted to the distribution and archive center for editorial review and incorporation into the information system.

2.3 CREDITS

To encourage individuals to compile data, clear ways must be developed to give credit to the individuals associated with data compilation, especially in multicomponent projects. Credit should be conveyed by citations that indicate the individuals responsible for the data in the dataset. The preferred citation, in a format similar to journal references, should be included in the metadata. Because datasets are sometimes combined for synthesis or modeling applications, a policy must be developed for citing data from multiple contributors.

In addition, individuals associated with data processing deserve credit for their involvement in the research process, involvement that often goes beyond providing a basic computer service.

2.4 DATA ENHANCEMENT

Even when well-documented datasets are available, they must be augmented, or enhanced, before they can be used in new applications. Enhancing the data, as used in the context of this paper, refers to data processing that is often required to perform integrated analyses and modeling within a multicomponent project. An enhanced dataset is a data product which is produced from an initial dataset and which is made complete and consistent within itself. The enhancement process addresses the practical aspects of incomplete datasets associated with field work (e.g., problems

A-04 4.tif

associated with broken instruments, miscalibrated sensors, observer errors, and unusual field events). Missing values or outliers may be replaced with estimated values with appropriate flags and documentation of the estimation process. Calibration factors may be adjusted on the basis of an analysis of the project data. Data may be aggregated or extrapolated to create datasets with appropriate spatial and temporal characteristics. Associated site data—such as data related to soils, topography, and climate—may be acquired from other existing sources as needed by the models and become part of the enhanced dataset.

2.5 DATA DISTRIBUTION

The Internet provides a powerful mechanism for facilitating the flow of information to a wide variety of users. During the active phase of a project, all of the individuals and teams and the project information system may have Web sites with links to each other. However, this technology does not automatically solve the concerns about documenting the data, integrating disperse data, or creating a long-term data archive. Although the technology for creating individual Web home pages is widely available, the functionality of the sites varies depending on the background and resources of the site administrators. The basic functions of a Web site that distributes data should include distributing data on multiple media, allowing data browsing even by users with limited Internet capabilities, maintaining and distributing metadata, offering a data-search-and-order capability, providing data security, and providing a data archive.

3.0 MODEL DESCRIPTION

The proposed information system model stresses providing a structure and incentives for individuals to prepare and release documented data. A five-tiered model is advocated: (1) individual investigators, (2) science teams, (3) a project information system, (4) a data distribution and archive center, and (5) a master data directory or catalog. Such a hybrid information system model builds on the organization and expertise of science teams (Tier 2) within the project. Functions may overlap or vary at each of the tiers, but the data flow must be defined and communications must be maintained among the tiers. This model contrasts with traditional information system models for research projects, which often do not address the long-term maintenance of data. The model is similar to that discussed by Strebel at al. (1994b), but with additional emphasis of the responsibilities of the science teams and the data distribution and archive center.

A-05 5.tif

3.1 INVESTIGATORS

Investigators help develop the overall project objectives and experimental plan, and each investigator is responsible for collecting, inputting, quality-assuring, processing, documenting, and eventually submitting their data to the project information system. Investigators are expected to follow general project guidelines and to meet a schedule for managing their data. Data preparation can be costly, often beyond the investigators' interests and resources. Therefore, it is essential not only to provide guidance and resources but also to ensure that individuals receive credit and that incentives are given, such as coauthorship of papers from participating science teams.

The investigators' roles may include collecting, inputting, and quality-assuring their data; documenting the data, following project guidelines; submitting data to science teams and/or the project information system; and authorizing final release of the data to others.

3.2 SCIENCE TEAMS

Most large field projects consist of several subgroups of investigators focused on studying different themes (e.g., an atmospheric team, a terrestrial team, a soils team) or investigators conducting similar work at different study areas or sites. In either case, these subgroups, or science teams, provide a level for addressing the issues of consistency and completeness in processing and documenting similar data. These teams are a logical level to bear the responsibility for defining common data collection methods, processing algorithms, and associated data formats; achieving data consistency and completeness; and meeting overall project data management schedules. Often these science teams publish papers as coauthors. Each science team would coordinate the development of data associated with its particular domain, including decisions about data requirements, data to be compiled, data formats, derived data products, and schedules. The team leader for each domain would provide technical leadership but may have additional resources and staff from the project information system to provide technical support. The project leader would provide oversight and coordination between the teams so that the data and data products from the teams would flow into the project information system.

The teams would be responsible for submitting clean and documented data to the project information system in the designated format or as close to it as possible. Team leaders would be responsible for collating data inputs from the investigators, performing quality assurance (QA) checks and metadata reviews, and possibly processing the collated data with standard algorithms. The QA checks include plots, statistics, and sorted lists; a review of logical combinations of variables;

A-06 6.tif

standardization of dates, times, coordinates, units, codes; and consistent data organization with consistent naming conventions. Comparing data from multiple sites can often facilitate QA checks that otherwise are not normally performed on a site-by-site basis.

Data management roles associated with the science teams may include establishing scientific objectives, methods, data requirements, and schedules for the team; maintaining communications among all of the project participants with respect to data needs and availability; collating investigator datasets, using common formats; conducting QA checks, standardizing data contents, and filling in missing or problem data as required; and submitting data to the project information system.

3.3 PROJECT INFORMATION SYSTEM

The project information system provides data processing and analysis support for the project investigators, distributes project information and data to project investigators, and performs other functions. The project information system works closely with investigators to prepare accurate and documented datasets by reviewing the QA checks, performing consistency and completeness checks, and reviewing documentation.

Although the project information system is integral to the project, some investigators view it as an unnecessary expense and an intrusion; therefore the flow of data to the project information system is often inhibited. To establish a linkage to expedite the flow of data at the later stages of the project, it is vital to involve the information system group in all aspects of the project from the beginning.

The project generally does not support distribution to a broader set of users or commit to long-term archiving of the data. As datasets are completed or at the end of the project, the group submits clean datasets and metadata to a distribution and archive center for distribution to the general public. The project may establish a Web home page to provide information to investigators within the project. The Web home pages for the project and the distribution and archive center would have links to each other.

Roles of the project information system may include gathering historical and background data; assisting in experimental plan development; establishing data and metadata standards for the project; assisting in the implementation of project infrastructure and field site operations; interacting with investigators and science teams; developing and operating the project information system; providing data analysis tools; processing auxiliary site, meteorological, and satellite/aircraft remote-sensing data; managing project data (back up, access control, etc.); maintaining an

A-07 7.tif

up-to-date inventory of data deliveries; maintaining and monitoring on-line data access; distributing copies of large off-line datasets; performing QA checks and processing data as requested; documenting and/or reviewing metadata; incorporating data in the project database; preparing final datasets for public access; and transferring data to a long-term data archive center.

Project information may include, but is not limited to, the following: project description; workshop and meeting summaries; project policies, guidelines, and standards; task checklists; investigator contact list; study area georeference information for each site; standard methods or algorithms; data inventory and data needs; models and model input requirements; geographic information system (GIS) base data; remote-sensing data; flight legs, videos, site photos, aerial photos, and maps; selected bibliographies of key papers; and extensive annotated bibliographies.

3.4 DISTRIBUTION AND ARCHIVE CENTER

The long-term distribution and archive center maintains the data, distributes data and metadata, and provides support to users of the data. The ORNL DAAC is an example of such a center. The center works with the project members to identify how best to serve their needs and to establish a schedule for transferring data. Data is transferred to the center as the project finishes or as individual datasets are completed. Data are reviewed for consistency and completeness and entered into the center's information management system. The center requires financial and institutional support to provide long-term archive of the data.

The distribution and archive center ensures data integrity during the transfer from data providers to the center; reviews or compiles documentation to ensure that it is accurate and complete and that it will provide present and future users (those not directly involved or familiar with the project) adequate information to understand the nature of the datasets and to assess the appropriateness of these for purposes beyond the scope of the original project; reviews or processes data to ensure that the data accurately represent the documentation and are consistent and complete; generates accurate and adequate metadata for use as search criteria that will enable users of varying scientific backgrounds to locate existing data to meet their needs; extracts valid keywords for search and order functions; acquires approval from data providers for final data product prior to public release; and loads the data and metadata into the data center information system.

The data center functions include providing data security with password protection and backups of data files; providing links to investigators, project science teams, and project information system Web pages; archiving data (multiple copies, long-term retention, migration to new medium as required); notifying users of any

A-08 8.tif

errors, changes, or updates to the data; advertising and actively marketing available data resources among scientists and other users; maintaining standard and valid keywords; supporting a data search-and-order capability; providing user support; and supporting data needs of decision makers, educators, and the general public.

The center will continue to maintain, distribute, and archive the data after the project finishes. The center supports the field project by providing a link between users and data providers. That is, the center generally may be able to answer many of the user questions, but if necessary the center could contact investigators for more information. In addition, to provide the post-project support, the center may routinely generate statistics on users of the data, solicit information from the investigators on publications and new data products, solicit information from the investigators and user community about potential errors and updates, provide a mechanism for investigators to exchange ideas on proposed activities or improvements, incorporate new data and data updates, and communicate changes and updates to users.

3.5 MASTER DATA DIRECTORY

Master data directories, or data catalogs, contain large collections of descriptions of datasets but do not work directly with or archive the data. The data directory contains pointers to the numerous locations that contain the data. Generally, directory-level metadata are very broad in their coverage so that they assist a scientist in learning what data are available for a particular area. The metadata allow a user to determine whether a dataset is interesting enough that additional specific information should be sought. Directory-level metadata can usually be used to reject a particular entry. However, to fully determine whether the dataset located in the directory is of real interest to the scientist, more information about that dataset is required at another level of the metadata hierarchy. To acquire the actual dataset described in the master data directory, the user must access the center or individual that maintains the dataset.

The NASA-funded GCMD is an example of a master data directory. A long-term goal of this effort is to provide science users with the ability to find and view information about science data regardless of what data system actually holds the metadata. The directory metadata interoperability has been addressed through two mechanisms: the development of a Directory Metadata Interchange Format (DIF) and the development of a DIF-based master directory that contains data across agencies, disciplines, and international boundaries. The DIF contains about 35 descriptive fields with controlled lists of keywords for many of the fields. The GCMD, including a description of the DIF structure, can be accessed at http://gcmd.gsfc.nasa.gov/.

A-09 9.tif

4.0 CONCLUSIONS

As a research community, we are challenged to develop and share the data resources generated by our research, both within projects and with external users. The sharing and integrating of data from multicomponent projects can be more efficient if several general principles are considered in developing a project information system. Development of the system should include establishing the flow of data from investigator to a long-term data distribution and archive center, maintaining communications among the several tiers involved in the flow, developing thematic science teams to address and perform data processing to ensure consistency and completeness of data for analysis and modeling needs, and instituting policies to give the data producers adequate credit for their efforts. In addition, to more fully share data, the scientific community should provide incentives for data sharing, recognize datasets as valuable research products, establish citation policy for data, establish guidelines for metadata, develop efficient data distribution systems (networks of inventories and archives), and promote the allocation of long-term financial and institutional support for scientific data.

5.0 REFERENCES

K.L. Gross, C.E. Pake, and the FLED Committee Members, *Future of Long-Term Ecological Data (FLED), Vol. I: Text of the Report*, Ecological Society of America, Washington, D.C., pp. 63-81, 1995.

W.K. Michener, J.W. Brunt, J. Helly, T.B. Kirchner, and S.G. Stafford, "Non-Geospatial Metadata for Ecology," *Ecological Applications*, in press.

National Research Council, *Finding the Forest in the Trees: The Challenge of Combining Diverse Environmental Data*, National Academy Press, Washington, D.C., pp. 3-11, 1995.

D.E. Strebel, D.R. Landis, K.F. Huemmrich, and B.W. Meeson, *Collected Data of the First ISLSCP Field Experiment, Vols. 1-5*, published on CD-ROM, National Aeronautics and Space Administration, Greenbelt, MD, 1994a.

D.E. Strebel, B.W. Meeson, and A.K. Nelson, "Scientific Information Systems: A Conceptual Framework." In *Environmental Information Management and Analysis: Ecosystem to Global Scales*, eds. W.K. Michener, J.W. Brunt, and S.G. Stafford, Taylor and Francis, Inc., Bristol, PA, pp. 59-85, 1994b.

A-10 10.tif